8. Data Sciences

Terms Used Throughout this Chapter

Application programming interface (API) is an interface between different parts of a computer programme to simplify the implementation and maintenance of software.

Artificial intelligence (AI) is the theory and development of computer systems able to perform tasks normally requiring human intelligence (such as visual perception or decision-making).

Bioinformatics is the science of collecting, analysing, and understanding complex biological data.

Cloud computing is the practice of using a network of remote servers hosted on the internet or in local data centres providing on-demand access to flexibly scalable resources to store, manage, and process data, rather than a local server or a personal computer.

Computational biology is the development and application of data-analytical and theoretical methods, mathematical modelling and computational simulation techniques to the study of biological systems.

Containers are standard units of software that package code and its dependencies. This enables the container to be shared between different computing environments to enable quick, reliable, and standardised analysis of data.

Data management is the development of architectures, policies, practices, and procedures to manage the data life cycle.

Data science is the science of using research methods, processes, algorithms, and systems to extract knowledge and insights from (typically large) structured and unstructured data.

Data science centre is used here as an umbrella term that includes data itself, its management, its bioinformatic processing and analysis (including but not limited to statistics, AI, and imaging) by scientists, the provision of code, tool and data services to internal users and the public, and links to (potential) local/national partners.

Data stewardship is a functional role in data management and governance, with responsibility for ensuring that data policies and standards turn into practice within the steward's domain. Data stewards support an organisation to leverage its data assets to full capacity.

FAIR Data Principles are a set of guiding principles for scientific data management and stewardship which aim to make data findable, accessible, interoperable and reusable.

Graphics processing unit (GPU) is a specialised electronic circuit designed to optimise the output of images to a display device. The highly parallel structure of GPUs makes them more efficient than general-purpose central processing units (CPUs) for algorithms that process large blocks of data in parallel.

High performance computing (HPC) is the practice of densely aggregating computing power, memory and storage at large scale linked via ultra-fast network links to deliver much higher performance than one could get out of a typical desktop computer or workstation in order to solve large problems in science, engineering, or business. It is used to solve computational problems that either are too large for standard computers or would take too long.

Machine learning is the study of algorithms that perform a specific task without explicit instructions. These algorithms automatically improve through experience, by relying on patterns and inferences in the sample data they are given.

Theory is a branch of science that uses mathematical models and abstractions of objects and systems to rationalise, explain, and predict natural phenomena. This is in contrast to experimental science, which uses experimental tools to probe these phenomena.

Background

The life sciences are being transformed by the presence of cheap and scalable technologies that can generate immense volumes of data, coupled with computational and infrastructural advancement that have radically changed the way scientists currently perform bioinformatics. Most biological subfields have recently experienced exponential increases in data, with biodata doubling in volume every 18 months. In addition to the increased use of nucleotide sequencing, biology is at the brink of an additional data revolution involving high throughput bioimaging technologies, such as those provisioned by the new EMBL Imaging Centre. Hence, the computational sciences in biology, including bioinformatics, modelling and computational biology, have become a necessary methodology which is relevant across many biological domains.

In this context, data science - the science of using large-scale data generation techniques along with tools and systems to extract insights from large structured and unstructured datasets - will be a key driver of the future of biology. Data science is also a major driver of startups and business innovations. There is the well-known adage that “data is the new oil”, with one of the implications being the sheer size and complexity of the IT infrastructure required. Novel analysis approaches, including those based on artificial intelligence (AI) and dealing with data of increased scale and complexity, will drive many biological innovations of the future.

Data science approaches in biology depend on coherent data structures, storage, and management, so that large datasets can be combined and utilised to fully exploit their potential for research. Consequently, the ability to manage, analyse, and make growing amounts of biodata accessible is of outstanding strategic importance for EMBL’s internal activities, as well as for external services and training provided to its member states. EMBL is uniquely positioned to lead in biological data science in Europe, since it combines capabilities and service facilities for generating large volumes of high-fidelity biodata, leading research activities in molecular biology and bioinformatics, and the hosting of widely used data repositories - all in a single institution.

The Opportunity

Molecular biology experiments lead to growing amounts of raw data that need storing with the appropriate metadata. This leads to various processing and analysis steps, sometimes in the context of complex international collaborations, and finally the need to make parts, or all, of the data available to the public, ideally in dedicated repositories. Similar to life sciences institutions everywhere, the various data flow scenarios at EMBL are complex, since data can have different origins and users, for example from within EMBL or from external researchers.

At EMBL, data ranges from: (i) Raw data generated by EMBL researchers (in the order of >8 Petabytes (PB) or >8 million Gigabytes annually); (ii) Raw data produced by external users of EMBL facilities and experimental services; (iii) Raw and processed data brought to EMBL via collaborators, consortia, or citizen science projects, as well as public and controlled-access data submitted by the scientific community to be hosted at the EMBL-EBI (now ingesting >10 PB of data annually); (iv) Several layers of processed data derived from (i)-(iii).

The large-scale data generation at EMBL is supported by several core facilities and structural biology, imaging, and genomics experimental services (Chapter 10: Scientific Services).

The growing volume and heterogeneity of the data have recently resulted in a massive growth of data science activities across EMBL’s sites. These have evolved rapidly, but in a rather independent manner with different solutions often being implemented across the organisation. The need to maintain efficiency and scientific

excellence in the context of growing dataset sizes; to maximise opportunities for collaboration on large-scale biodata sets within EMBL; to foster EMBL's future engagement in cross-disciplinary large-scale research efforts, especially in the new directions in this Programme; and the need to enable optimal interactions with member state researchers and research infrastructures requires **putting biodata at the centre of our research strategy.**

As EMBL undertakes new scientific directions under the Molecular to Ecosystems Programme, a highly coordinated approach to biodata will be essential to keep EMBL at the forefront of molecular biology; to allow it to efficiently and economically deal with large-scale data; to attract and develop young talents and foster careers; to make EMBL ready to embed the most modern analytical approaches in big data and AI; and to create new bridges between disciplines. As such, EMBL can act as a role model to empower the development of biological data sciences across Europe.

Aims

Throughout the next EMBL Programme, EMBL plans to establish and grow a **data science centre** which will coordinate biodata activities across EMBL, allowing for flexible and adaptable management. The name data science centre is used as an umbrella term that includes data itself, its management and bioinformatic processing by dedicated staff, the provision of code, tools and data services to internal users and the public, and links to (potential) local or national partners, as well as the promotion of open science principles.

EMBL's data science centre will be highly coordinated, since needs at each site are heterogeneous and complex. Support staff and teams relevant to EMBL's data science centre can therefore be spread across the EMBL sites. Overall it is expected that these factors will lead to higher quality, efficient, and robust internal and external services, crucial to EMBL.

EMBL's data science centre will also leverage EMBL's data science capabilities to foster and integrate data-driven biology and data science research in Europe through mutually beneficial collaborations. It will also allow EMBL to maximise participation in internationally and nationally funded activities relevant to the biological data sciences aligned with EMBL's missions in science and infrastructures.

EMBL's data science centre may take on some or all of the following roles:

1. **Provide internal research support for biologists**: Data science research support teams will provide research support not only in classical bioinformatics domains, but also in omics analysis, statistics, machine learning, visualisation, bioimage analysis, computer-simulation based experimental design and modelling. They will additionally be associated with internal experts (group and team leaders) in the respective areas relevant to data science.

2. **Facilitate research advances in data science**: While the main role of the data science centre is to enable, there will also be an important place for performing research to develop and improve methods and tools, and for further discovery science. Thus the data science centre can have embedded group and team leaders, and attract increased capacity for innovative technological solutions via recruitment of skilled data science professionals, postdocs, and PhD students.

3. **Develop common representation of data, conventions, workflows**: The data science centre will enable researchers to capture increasing amounts of metadata, in order to maximise the utility of large-scale biodata sets. This will also enable EMBL-EBI best practice to be propagated for simpler and more coherent data management, to transfer to national and global depositories, and to make data findable, accessible, interoperable, and reusable (FAIR Data Principles; https://www.go-fair.org/fair-principles).

4. **Provide training and career development**: The data science centre will further deliver training to develop the careers of EMBL's researchers as well as member state researchers. EMBL is in a fortunate situation to have many in-house specialists who can deliver, in a joint effort, a wide spectrum of data science training, from the basics of data analysis, statistics and bioinformatics to advanced and newly arising topics including AI and theoretical biology.

5. **Offer critical public data resources**: EMBL can provide services such as public data resources and tools, for the growing volumes of biodata and the diversity of biological applications, to empower the scientific community to work with big datasets. EMBL will also implement and promote open biodata practices, by making visible and enabling FAIR data and tools.

6. **Act as bridges to member state activities**: The data science centre at all EMBL sites will have the ability to get involved in national infrastructures and member state activities. The data science centre will interact with national ELIXIR nodes, and engage in collaborative fundraising (public and industry funds) for external facing bioinformatics services and training beneficial to both the EMBL site and the member states.

EMBL's data science centre will be the primary arena to fulfil EMBL's commitment to **open science** principles and practices. Open science includes open data and open methods, with one aspect being the FAIR guiding principles for scientific data management and stewardship, which intend to improve the findability, accessibility, interoperability, and reusability of data. In spite of its clear advantages, open science in biology is still in its relatively early days. Currently, only positive and fully interpreted results tend to be published by biological communities. By comparison, raw data as well as protocols and software used to process these data are often not published, which affects reusability and reproducibility of science. Thus the practical application of open science principles will require both policy development and implementation planning. EMBL aims to act as a leader by using and promoting open science practices. These practices encompass the management of research software and data products, the data that streams from community use of the research facilities at EMBL, and the wider, global contribution to data management via the public data resources hosted by EMBL on behalf of the research community.

The data science centre will also implement considerations around data protection, including EMBL's implementation of the EU's General Data Protection Regulation (GDPR). Research activities with human data will significantly increase in the future, including controlled access data types. Thus, data protection and all of its implications will need more attention. Here, EMBL will continue to develop appropriate policies and processes which can facilitate robust collaborations between EMBL and member state researchers in a shared legal and regulatory framework.

EMBL's Approach

The implementation of the data science centre requires careful and coordinated planning. EMBL has identified five priority areas that will be the focus for data sciences throughout the next EMBL Programme. These priority areas will enable EMBL to maximise the discovery of research data, to effectively manage the growing volumes of biomolecular data, and to engage with the diversity of biological applications and the varied needs of a growing and global scientific community (Figure DS1).

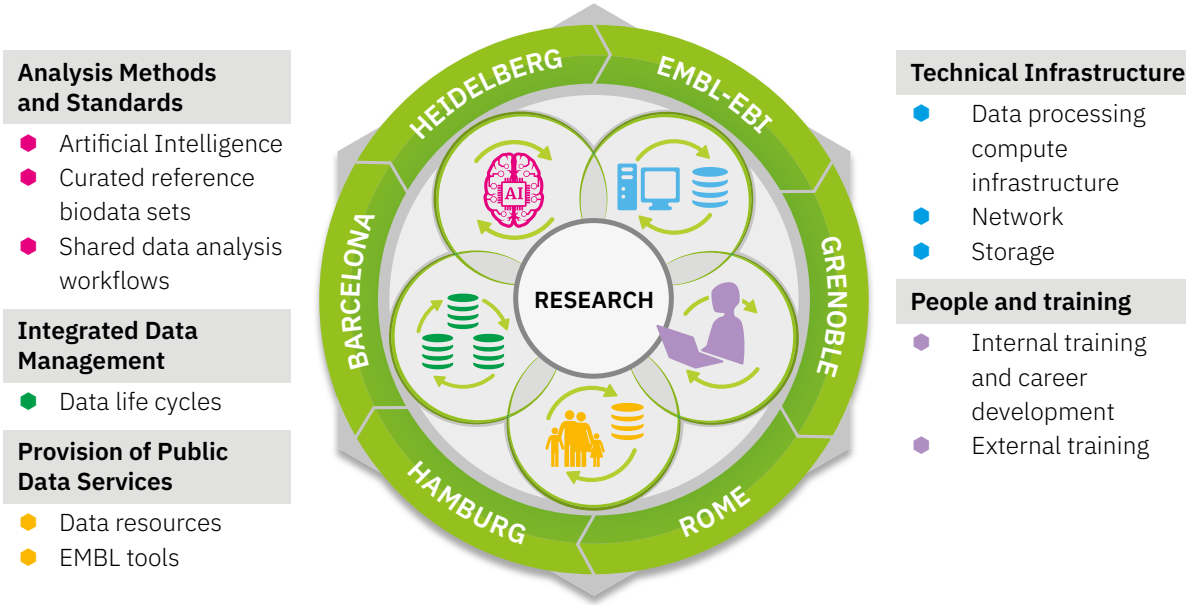

Figure DS1 | Data Sciences Programme, connecting data science at all EMBL sites.
EMBL's new Data Sciences Programme aims to connect data science focused on five priority areas at all EMBL sites. This activity will be highly connected across sites to leverage synergies across all of EMBL and to share expertise and experience while integrating into the local research and service landscape to ensure bespoke solutions where needed.

Analysis Methods and Standards

Artificial Intelligence (AI)

AI technologies will be key drivers of innovation in data science in the coming years. EMBL's research and service teams are already active in various aspects of AI, both as users and developers of novel AI approaches. This includes bioimage analysis; investigating human genetic variation in non-coding regulatory DNA; and integrating somatic mutation and pathology image data for cancer classification. AI is an area of expected future growth, since AI approaches such as deep learning have ample potential to facilitate the analysis and integration of structured and unstructured data, especially in the context of large-scale datasets. EMBL is well-positioned to become a leading player at the interface of AI research and molecular biology, where

EMBL's strengths include:

- The ability to manage and integrate large amounts of biodata data types (e.g. omics, bioimaging);
- Expertise in curating data (e.g. to label biodata with the "ground-truth");
- Data-producing service facilities and involvement in pan-European consortium activities creating large-scale reference datasets which are highly relevant for training AI software;
- *In silico* AI research to build and leverage AI methods in the life sciences.

EMBL will continue to monitor progress in the general machine learning and AI research communities and adapt relevant approaches to biological data or develop new ones where appropriate. EMBL's aims in this are threefold: **driving biological discovery science, contributing to machine learning/AI research, and improving EMBL's services**. One example of current research at EMBL is an approach to deep transfer learning connecting histopathology and patterns of mutation and gene expression in cancer (Chapter 6: Human Ecosystems), which illustrates EMBL's engagement in leading-edge AI developments and applications in a research setting.

As part of the new EMBL programme, EMBL will engage in AI research in several areas particularly relevant to biology, including:

- **Un-/semi-/self-supervised learning to tackle lack of training data**. This aspect will be especially relevant to cell type classification and disease classification problems, where ground truth data are often lacking.
- **Explainability and uncertainty quantification**. AI applications in the life sciences can be empowered by obtaining insights into the relevant AI models. Deriving causal and/or mechanistic insights, and quantifying model uncertainty, will thus represent a focus area.
- **Interpretable low-dimensional representations and metric learning**. Complex multimodal high-dimensional data (such as single-cell multi omics data) require methods for dimensionality reduction and analysis of interrelations, which will represent another focus of EMBL's AI research.
- **Privacy-aware and federated learning**. In the context of human-derived data, federated algorithms and infrastructures that can preserve privacy while allowing aggregate information across different compute centers and jurisdictions will be developed.

The conceptual underpinnings of AI methodology in the data sciences theme will be tightly linked to the new theory theme, which addresses questions such as model conceptualisation, interpretation, and critique (Chapter 9: Theory at EMBL).

To advance machine learning in the life sciences, EMBL will focus on **selected areas** where EMBL is generating data or managing unique data resources. Strategically, EMBL sees great potential added value of AI to biological research and services in the annotation and interpretation of sequences (genomes, transcripts, and proteins) and sequence variants, bioimages (segmentation, navigation, connectomics, and phenotyping), and biomolecular structures. To achieve these aims, the service aspects around data science will facilitate the annotation, navigation, cross-modal integration, and interpretation of huge amounts of structural, bio-image, and omics data using modern algorithms and AI. This includes the ability to train and compute directly on archived data, such as the BioImage Archive recently established by EMBL (Chapter 10: Scientific Services). It also includes the need to embed advanced AI algorithms directly into research and service infrastructures that generate these data, thus facilitating their use by EMBL researchers and users of EMBL's external services.

Key requirements for assessing progress in AI are benchmark datasets and open challenges, where tools are objectively compared on representative datasets. Critically, there is a scarcity of such reference datasets for major biodata domains, including microscopy, where few reference datasets exist. EMBL researchers will take a leading role in **establishing, collecting, and disseminating curated reference biodata sets** and problems, for which AI-based methods are particularly suitable. EMBL researchers will also develop baseline solutions and the required infrastructure components, which will allow biologists and AI researchers to synergise and address open questions in molecular biology.

To bolster AI efforts, EMBL aims to engage with **AI research communities** at multiple levels to foster developments inside and outside of EMBL. AI and machine learning research at EMBL will be connected to European networks and AI initiatives, including the European Lab for Learning & Intelligent Systems (ELLIS; https://ellis.eu). In this context, EMBL's successful proposal to create ELLIS Life, a cross-institutional ELLIS unit between EMBL, the German Cancer Research Center (DKFZ), and the University of Heidelberg, will foster this interaction.

A key advantage of AI methods is that they can be reused for a wide range of recurring tasks, where biologists only need to provide input or training data to parametrise an existing algorithm. Alongside EMBL's role within ELLIS, EMBL has leadership roles in several AI communities which distribute pre-trained models for such recurring tasks via repositories of code and parameters, so-called model zoos, such as the model zoo for genomics (https://kipoi.org) and the model zoo for bioimaging models which is expected to be launched in late 2020. While these repositories are important in the core AI domain to ensure reproducibility and full exploitation of available training data, they become even more crucial in domains where the models are deployed by non-experts (e.g. wet lab biologists) as they provide a level of trust in the choice of algorithms.

EMBL also aims to strategically **hire more AI talent** to further strengthen research at the interface of AI and molecular biology. This includes recruiting researchers and service staff already trained in the mathematical, statistical, and computer science foundations of machine learning, and training them in biology as well as training biologists in the advanced use and adaptation of machine learning/AI algorithms. It is foreseen that innovations in AI are likely to strengthen and help drive EMBL's future biological discovery research, as well as innovations enabling technology transfer, and service provision. Strengthening AI research in Europe could counteract the brain drain of AI talent to North America and ultimately raise the competitiveness in "biodata AI" in EMBL's member states. The tools and resources developed by EMBL, reference datasets archived at the EMBL-EBI, and cloud resources with access to graphics processing units (GPUs) and data, will facilitate wider adoption of AI in biology in Europe.

Shared Data Analysis Workflows

The growing utility of data across all biological domains implies a need for more efficient data science practices and economic use of advanced computing capabilities. EMBL's past success has partly been grounded in its fast-moving, "bottom-up", decentralised planning of research studies. However, this brings with it the risk of duplication and redundancy in the development of data analytic workflows. Large-scale consortia studies such as Pan-Cancer Analysis of Whole Genomes (PCAWG) or the Human Cell Atlas (HCA) have highlighted key advantages of common (shared) computational workflows for recurring tasks (Figure DS2). For example, shared workflows can save productive time, increase reproducibility and scalability, and thus can free resources to tackle new questions.

To support EMBL's research including the new research themes, EMBL aims to extend the use of such 'standardised' workflows for recurrent activities, such as what is commonplace at the Broad Institute in the United States of America. This will be first achieved within EMBL, to subsequently enable and empower

external research communities across EMBL's member states. These harmonisation efforts will embrace the relevant community standards like those developed by the Global Alliance for Genomics and Health (GA4GH) or the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI) to maximise international comparability and reproducibility.

Software packaging, deployment using classical high performing computing (HPCs) and clouds, and life cycle management services (e.g. Bioconductor) and containerisation (e.g. Docker, Singularity) will facilitate workflow dissemination. The development of such workflows will also provide additional opportunities for EMBL to lead in coordinating the development of standards within relevant biological communities.

Figure DS2 | Shareable workflows to enable standardised analysis of protected data.
Shareable analysis tools (containers) can be sent to locations where large data sets reside, such as a particular EMBL site, or, in the case of human data research protected data hubs, as previously achieved in the Pan Cancer Project led by EMBL (www.nature.com/collections/PCAWG), to enable the standardised analysis of biodata irrespective of where the data are stored.

Integrated Research Data Management

EMBL's approach to coordinated data sciences aims for seamless data flow throughout the whole life cycle of data, from sample to production of the raw data to data analysis and finally to open access publication of the findings and deposition of the data in the appropriate databases (Figure DS3). This will not only lead to all EMBL data being made FAIR, but higher-quality data will be more efficiently produced and analysed, as well as allow for EMBL research outputs to be more rapidly made available for scientists in the member states and beyond.

Data life cycles are often complex involving individual and highly customised workflows according to the needs of the specific discipline. But general principles are shared across disciplines, including workflow stages such as data processing and analysis, quality assurance, integration with other data, dealing with sensitive data where access needs to be controlled, sharing within or across international consortia, data archiving, and publishing. Currently, data intensive research projects typically involve a lot of data movement, and thus, copying and duplication across IT infrastructures, different data storage devices, clouds, and data archiving systems can be made much more efficient.

Implementation of optimal data management policies, processes and tools will require expertise in data management planning and data stewardship to integrate new efforts with those that already exist at EMBL such as the data management applications available for the microscopy core facilities. Researchers, facilities managers, IT staff, and public data resource managers will all be involved in implementing a software framework spanning from wet lab groups all the way to the open databases at EMBL-EBI. A new internal data deposition brokering network, helping the process from locally generated data towards public deposition and representation, will be hosted within the data science centre. Dedicated data scientists will also support wet lab researchers through providing services in data analysis, integration, and FAIR data management and sharing. These data scientists will facilitate good practice in scientific computing and enable community building across EMBL sites.

EMBL's data management tools will in the future:

- Permanently link individual datasets through stable identifiers such as DOIs or accession numbers (Figure DS3, dashed blue arrow), allowing data to be consistently referred to within electronic lab notebooks and, at the same, avoiding data duplication;
- Aid collaborative efforts across EMBL by giving a comprehensive overview of where data are stored and organised according to a project, and on their current state of processing (eg. quality assurance, analysis, publication, and archiving);
- Manage access to and open the sharing of data to facilitate data integration across data sources, and to maximise the quantity, quality, and reusability of research outputs;
- Support key functions needed for physical data movement including long-term archiving with the inclusion of relevant metadata;
- Help EMBL's data-generating core and service facilities to adapt their IT workflows, increasing their efficiency and quality.

Pilot implementation initiatives are in progress and focus on multi-omics data, X-ray diffraction data (Figure DS3), and imaging data. These pilot initiatives will help find answers to big data questions and establish use cases for shared use of IT resources and data analysis and management across EMBL. Member state users of these EMBL facilities and services will likewise benefit from enhanced data management, data representation, and data flow capabilities.

The planned initiatives around research data management could have a significant impact beyond EMBL. Projects like Euro-BioImaging and Instruct-ERIC (Chapter 13: Integrating European Life Science) may be starting points for scaling beyond EMBL as the problem is common to all research institutions and particularly those supporting life sciences facilities and services. Efforts could also be linked to the ELIXIR-led EOSC-Life initiative aiming to provide generic services accessible across Europe to support open science within the context of the EOSC (Chapter 13: Integrating European Life Science). Through these activities EMBL could help develop open science best practice that will be at the disposal of the wider scientific community to adopt and use, or directly provide data science services and infrastructures to member states.

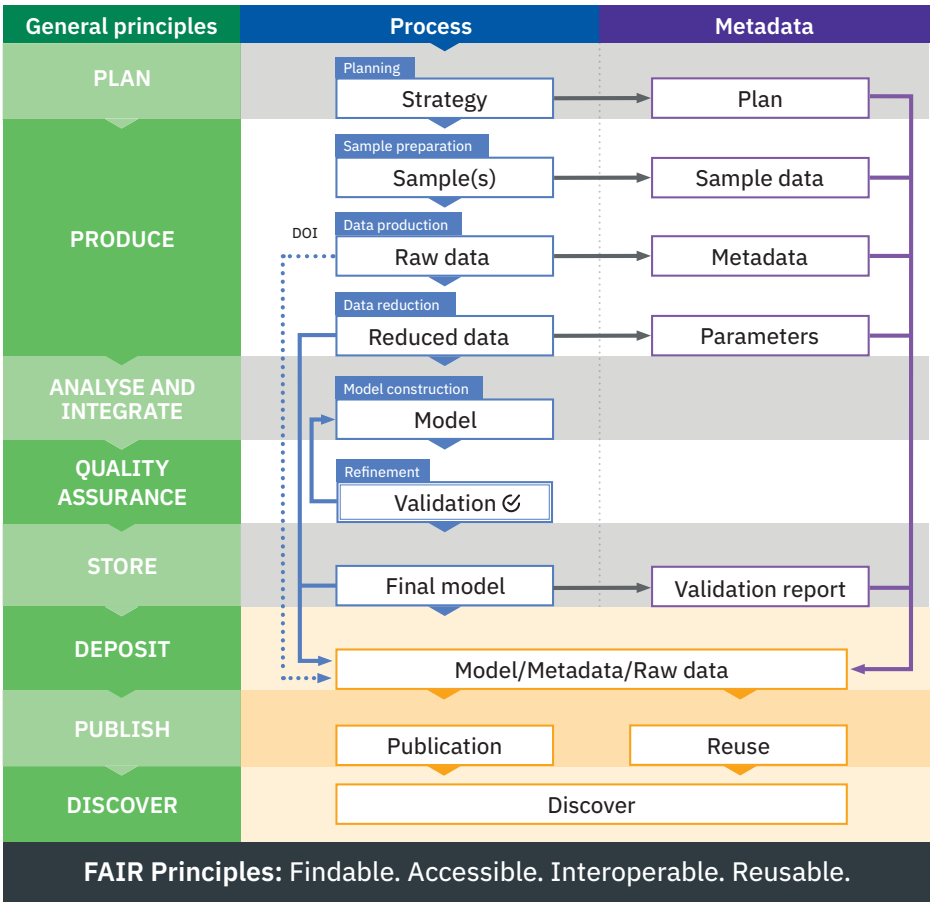

Figure DS3 | EMBL data management workflow.
Supported by FAIR principles, the general principles of the EMBL data management workflow guide the flow of data and metadata for bespoke research examples.

Provision of Public Data Services

Molecular Data Resources

Through EMBL-EBI, EMBL has a community mandate to develop and maintain a comprehensive array of open access and up-to-date **molecular data resources** as a key part of its services mission (Chapter 10: Scientific Services). These data resources cover a huge range of molecular biology subdisciplines including nucleotide sequence data, bioimages, protein sequences and families, chemical biology, structural biology, systems, pathways, ontologies, and scientific literature. In short, EMBL-EBI's data resources collate, integrate, curate, and make the world's scientific biodata freely available to all.

The EMBL-EBI data resources include deposition databases (archives) that store primary experimental data generated by scientists worldwide, as well as knowledge bases that integrate and add value to experimental data, making it easier to use and understand, with many having both functions. Through EMBL's engagement with the research community - whether that be the development of data resources to meet continuously changing research needs or via training and industry engagement programmes - EMBL strongly supports and propagates the concept of FAIR biodata.

EMBL-EBI data services are managed as an institutional commitment on behalf of the research community, rather than as the sole responsibility of individual group or team leaders. These open and searchable resources provide all researchers with direct access to the scientific record, enable access to and reuse

of experimental data to verify original results, and, by combining multiple data records, provide analytical insights. Deposition databases also provide reference data for the research community. Through the use of search tools, researchers can rapidly compare their own unpublished data with open access datasets.

EMBL-EBI monitors the usage of resources (for example, web hits, data downloads, citations), which feeds into processes for managing the data resources from start-up, to full operation, to, in some cases, retirement. The management of these resources is driven by scientific need. Therefore, in the same way that new resources may be launched to meet emerging needs (such as the BioImage Archive), others may be downgraded as needs for particular technologies or data types change. The data stewardship expertise at the institutional level at EMBL-EBI is of fundamental importance in managing these processes.

Research Software

Research software produced at EMBL is of two main types: (i) data analysis workflows or "scripts" associated with a specific dataset, scientific question and, often, publication; and (ii) more generic methods, algorithms, or tools intended to be useful across a range of applications.

Many computational tools developed by EMBL researchers are highly used by the scientific community. Some of the highly used tools in 2019 were InterProScan (accessed 55 million times to annotate sequences), DESeq2 (downloaded 286,000 times for differential gene expression analysis), Ensembl Variant Effect Predictor (used 177,000 times to analyse over 180 million variants), ilastik (downloaded 60,000 times for image classification and segmentation), webPRANK (a phylogeny aware multiple-sequence aligner used 10,000 times), and limix (downloaded 8,000 times for genetic analysis of multiple traits).

Currently, research tools developed by EMBL researchers are made publicly available according to the needs of individual groups. This is dependent on the availability of funding, resources, and time, rather than being coordinated and managed institutionally. The majority of public data resources and tools are currently at EMBL-EBI, and a growing number of such data services have recently emerged from other EMBL sites. In addition, the large-scale provision of services to produce bioimage data by EMBL researchers and external users within the new EMBL Imaging Centre will create immediate needs in image data analysis in the user community from member states. EMBL aims to extend its expertise in bioimage data analysis to develop new image data analysis methods as well as provide these tools in a user-friendly manner to the internal and external user community via robust services.

In the future, EMBL will strategically align the provision and life cycle management of external facing data resources and tools declared to be a service at all EMBL sites. EMBL will develop a more cohesive framework for deciding which research data and tools should be supported institutionally and how those resources should be developed and supported both in terms of skills and expertise and technical infrastructure.

The institutional approach to public data services provision aims to:

- Develop processes and support to identify, develop and maintain computational research tools and data resources that can be declared to be of enough strategic value or impact to be supported by EMBL as a service;
- Have clear rules for open vs. closed source, defined in view of the competition of algorithms and commercialisation of software, with a strong preference towards open software;
- Train people involved in services in software engineering to achieve robust, production-level tools and software life cycle management;

- Ensure that the tools and data resources deemed to be of value as a service will from then on share a common brand or certification, which EMBL expects to result in the increased visibility of tools or data resources, both within and outside of EMBL;
- Ensure that tools and resources have a path to be officially retired, for example if the relevant technology becomes obsolete.

As EMBL embarks on new scientific directions, it is envisioned that EMBL can pioneer the provision of public data services (both data resources and tools) for these new initiatives, as well as provide data platforms for new communities of users such as ecologists, marine biologists, clinicians, and epidemiologists. This can be similar to how EMBL currently provides bioinformatics resources to clinicians and pharmaceutical researchers for genomic medicine diagnosis and drug discovery.

Technical Infrastructure

Compute requirements, including for classical high performance, high throughput and cloud computing will grow with EMBL's needs for data generation and analysis. These call for broader sharing of access to IT resources, including improved distributed access to HPC, cloud services, and long-term data storage at all EMBL sites.

Cloud technology has the vast potential to simplify connecting and federating distributed IT infrastructures, as well as enable cross-site research studies, where data may not always be able to move between EMBL sites at the generation and analysis stages (Figure DS2). Currently EMBL-EBI operates the Embassy Cloud which, when using OpenStack, enables fast and scalable provisioning of virtual machines in the manner of an infrastructure-as-a-service (IaaS) cloud solution. Thus, this allows for dissemination of tools and computation on big data for local scientists and remote collaborators alike. Similar benefits from reuse of pre-packaged or "containerised" workflows and tools are provided by the container hosting platforms operated both at EMBL-EBI and EMBL Heidelberg. The concept of an EMBL Science Cloud has recently been launched to promote the use of clouds, initially building on existing infrastructures such as the EMBL-EBI Embassy Cloud and the 3D Cloud in Heidelberg (Chapter 14: People, Processes, and Places).

For the next EMBL Programme, EMBL will increase the accessibility of cloud resources (e.g. OpenStack-based) and the use of cloud application programming interface (APIs), augmented by software for orchestration (e.g. Kubernetes), to create a distributed computing environment enabling data processing with similar tools, irrespective of where the data are. This is driven by the fact that more and more data cannot easily flow to, from, or between EMBL sites owing to their size or legal or regulatory constraints. This could include datasets generated from EMBL's new directions such as Planetary Biology and Human Ecosystems. EMBL internal clouds also need to be able to interact with commercial clouds, in a hybrid cloud model, wherever this is a better solution, whether for financial, legal, or regulatory reasons. The increased usage of AI at EMBL will also result in future GPU needs that EMBL aims to satisfy.

With regard to data storage, all stages in the life cycle of a particular dataset will be appropriately tracked as described under **Integrated Research Data Management**, including the context of the experiment or project, the physical location, copies, and ownership. **Modern data storage solutions**, such as object storage to complement file system storage, will be explored to support this in a cost-efficient manner. The FIRE service established at Hinxton provides a scalable resilient data storage infrastructure that can ingest over 1PB of data a month and has scaled already to over 30 petabytes.

People and Training

Data science is a multi-disciplinary area which includes bioinformaticians, statisticians, mathematicians, and theoreticians. Personnel involved in EMBL's data science activities are based at all of EMBL sites and can fall into a number of different categories (Figure DS4).

In silico researchers in purely dry groups make up roughly 18% of all research groups across EMBL and are located at all EMBL sites, with the majority residing at EMBL-EBI and EMBL Heidelberg. Research staff undertaking data sciences in predominantly wet labs and core facilities (33%) can range from dedicated bioinformaticians with bioinformatics and/or data analysis expertise in a specific domain to experimental scientists working at the wet-dry boundary. A growing number of wet lab scientists are acquiring data science skills to become "hybrid" (wet-dry) researchers to analyse individually generated large-scale data. A few dedicated individuals (~10) provide primarily internal bioinformatics and data analysis core services, as well as specialist expertise and advice for EMBL wet lab scientists. They also provide some internal bioinformatics and data science training, often supported by data scientists within the research groups. Staff responsible for EMBL's public bioinformatics service activities make up roughly 400 of the 1,800 EMBL personnel. Although now mostly at EMBL-EBI, as the number of outward facing services increases at the other EMBL sites, these capacities will expand during the new Programme.

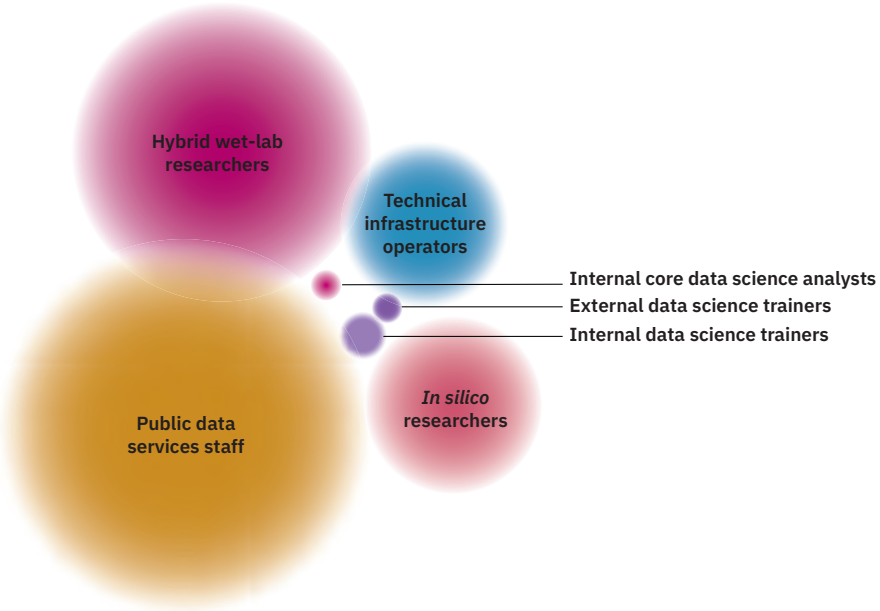

Figure DS4 | Proportional representation of the number of data scientists by their main activity.

Data scientists at EMBL can fall into multiple categories spanning *in silico* researchers, a growing proportion of hybrid researchers, provision of internal and external bioinformatics training, as well as the provision of infrastructure and technical services.

The data science centre needs to ensure that staff in all of the categories listed are adequately and centrally supported to carry out their respective functions in this competitive and fast-evolving area. As such, EMBL needs to ensure it is positioned to identify and attract the **best talent** in strategically important research areas; ensure career data scientists are **recognised** for their contributions; provide mechanisms for **networking** with bioinformatics peers across EMBL and externally; and ensure appropriate **workload management**, especially for those who provide core and public data and analysis services.

The need for central support for **internal training and career development** (Chapter 14: People, Processes, and Places; Career Development) for the new generation of data scientists at the frontier of wet and dry biology is expected to grow. Future training will enable EMBL staff at all levels to stay current in their knowledge and use of data science methods and tools (including for new approaches such as AI) and prepare them for their future careers. Access to data management training will allow EMBL researchers to efficiently use data management resources on offer as well as comply with data management strategies and policies.

As the data science centre develops, EMBL anticipates that staff will become **key providers of external training**. EMBL-EBI's Train Online platform, which already incorporates some courses on basic data management principles, can be used to support advanced, face-to-face external training at EMBL and in the member states (Chapter 11: Training). Ongoing participation in ELIXIR's training platform plus the creation of pop-up training rooms, most likely using commercial cloud-based compute, will be developed to cost-effectively manage the 'bursty' need for compute to support training in and beyond the member states. Empowering scientists to construct their own solutions will be one pillar of these new directions, which will ensure that the very best of EMBL data science is shared with the member states.

Impact

EMBL envisions that a new EMBL-wide approach to data sciences could have significant impact externally including in EMBL's member states and the global scientific community, as the problem of growing datasets is common to all research institutions in the life sciences. Specifically, the EMBL-wide approach to data sciences will lead to:

Research support – Provide enhanced research support for member state users of EMBL's core and service facilities in managing, analysing, sharing, and archiving their data. Member state researchers will not only benefit from a more streamlined data flow, but will also be able to reuse shared standardised data processing workflows developed at EMBL to accelerate their research and improve the reproducibility of their analyses.

Training and career advancement – Help develop the careers of wet lab biologists and data scientists within Europe's increasingly interdisciplinary life sciences research landscape. In addition to offering external courses at the six EMBL sites, EMBL will work towards developing a network of partners in other member states to deliver data science training locally.

Promoting research collaborations with member states – Provision EMBL's enabling data science resources and tools for data-intensive collaborations involving member state scientists, such as the European Open Science Cloud (EOSC), the Human Cell Atlas (HCA), as well as the initiative 1+ Million Genomes which works towards access to at least 1 million genomes in the EU by 2022.

Formation and strengthening of local hubs with national ELIXIR nodes – Engagement of individual EMBL sites with member states.

Support in open science – Support member state scientists in producing FAIR data, to facilitate their adherence to grant requirements and (inter)national standards.

Through these efforts EMBL could serve as a role model for life sciences institutions that need to cope with growing volumes of data, or even directly provide infrastructures to member states. This would particularly empower those member state users that in their home institutions lack the facilities and infrastructures to deal with storing and analysing large-scale datasets.